韓國컴퓨터情報學會 論文誌
第14卷 第11號, 2009. 11.

2009-14-11-5-2

공식통계의 추론통제 전략
- 정부의 특허경비지원사업 사례를 중심으로 -

이 덕 성*, 최 인 수**

A Strategy for Inference Control of Official Statistics
- Centering around the Patent Application Expense Support Project -

Duck-Sung Lee*, In-Soo Choi**

요 약

정부와 지역사회를 위해 나라에서는 공식통계를 수집하는데, 이러한 공식통계는 정부 정책이나 프로그램의 유효성을 평가하는 데에도 사용된다. 따라서 공식통계는 정확한 사실을 바탕으로 수집되고 공표되어야 한다고 본다. 정확하지 못한 공식통계는 정부 정책이나 프로그램의 평가를 그르치게 하기 때문이다. 오늘날 여러 통계기관이 주가되는 공식통계 전달 매체로서 집계기능을 발휘하는 OLAP 데이터 큐브를 채택하고 있는데, 이러한 데이터 큐브에서의 기밀을 보호하는 것도 아주 중요한 문제로 대두되고 있다. 왜냐하면 데이터 큐브가 악의적 추론을 당하게 되면 데이터 큐브에서 기밀유지를 해야 할 중요부분이 누설될 수 있기 때문이다. 저자들은 먼저 정확한 큐브를 작성하게 하고 큐브에서의 기밀누설을 막을 수 있는 OLAP 데이터 큐브에서의 추론통제 프로세스를 제안한 바 있다. 본 연구에서는 이 추론통제 프로세스를 사용하여 공식통계의 추론통제 전략을 수립하는 것을 목적으로 하고 있으며, 정부의 특허경비지원사업을 사례로 삼고 있다.

Abstract

Official statistics which are collected for governments and the community can be used to assess the effectiveness of governments' policies and programs. Thus, official statistics should be collected and presented based on correct findings. Erroneous official statistics will lead to lower quality results in assessing those policies and programs. Many statistical agencies, today, use on-line analytical processing (OLAP) data cubes which support OLAP tasks like aggregation and subtotals as a key part of their dissemination strategy of official statistics. Confidentiality protection in data cubes also should be made. However, sensitive parts of data cubes including micro data may be

• 제1저자 : 이덕성 교신저자 : 최인수
• 투고일 : 2009. 10. 19, 심사일 : 2009. 10. 28, 게재확정일 : 2009. 11. 26.
* 숭실대학교 대학원 산업·정보시스템공학과 재학 ** 숭실대학교 산업·정보시스템공학과 교수
※ 본 연구는 숭실대학교 교내 연구비 지원으로 이루어졌음.

disclosed by malicious inferences. The authors have suggested an inference control process in OLAP data cubes which preventing erroneous cube creating and securing cubes against privacy breaches. The objective of this study is to establish a strategy for inference control of official statistics using the inference control process by taking the case of the Patent Application Expense Support Project.

▸ Keyword : 데이터 큐브(Data Cube), 추론통제(Inference Control), 기밀보호(Confidentiality Protection), 마이크로 데이터(Micro Data), 공식통계(Official Statistics)

Ⅰ. 서 론

어느 나라를 막론하고 통계를 가장 많이 이용하는 곳도 정부이며, 통계를 가장 많이 생성하는 곳도 정부라고 한다. 복잡하고 방대한 행정업무를 수행하기 위해서, 장래에 쓰일 정책목적을 결정하기 위해서, 그리고 일정기간동안 추진할 경제사회 발전계획을 수립하고 시행하기 위해서 정부는 통계를 생성하고 이용하는데, 이러한 정부의 목적이 제대로 성취되기 위해서는 무엇보다도 먼저 통계정보 자체가 국가적 차원에서 본 정확한 정보가 되어야 한다고 본다. 즉, 통계정보가 왜곡되지 않아야 한다는 뜻이다. 그리고 국가의 목표가 달성되고 진전되는 상황을 국민에게 홍보하기 위해서 정부는 통계를 생성하고 이용하기도 하는데, 이러한 정부의 목적을 원활히 달성하자면 UN의 통계 기밀보호 원칙[1]을 지키는 한도 내에서 최대한 많은 양의 통계정보가 국민에게 공개되어야 한다고 본다.

우리나라에서도 국정을 운영하는데 필요한 국가적 단위의 통계(official statistics; 이하 공식통계로 기재)를 대부분 생성하고 있다. 이러한 공식통계는 정부정책의 기본 인프라로서의 역할을 하게 된다. 즉, 작성된 통계수치에 의해 정책입안과 시행이 되기 때문에 국민생활에 직접적인 영향을 미치는 것이 공식통계라는 뜻이다. 그런데 이러한 공식통계가 왜곡되어 심각한 문제를 야기 시키는 경우가 종종 있다. 예로 정부기관이나 민간기관이 작성하는 통계의 경우 기관의 업적을 과대포장해서 홍보용으로 이용하거나 잘못된 행정을 감추거나 호도하는데 이용하는 등 통계가 왜곡되기도 한다. 또한 개념정의의 차이로 인해 왜곡되기도 한다. 실업률통계의 경우 체감실업률과 국가통계 실업률 수치에 차이가 많이 생겨 혼란이 생기기도 하는데, 이는 ILO와 OECD의 실업정의가 다른데서 기인하는 것으로 본다. 공식통계의 왜곡으로 인해 국가와 국민이 입는 손실이 막대하다는 것은 주지의 사실이다[2].

개인이나 가족, 비즈니스 같은 것에 관련된 통계용으로 국가 통계청이 직접 수집하거나 행정기관 같은 곳에서 얻는 데이터를 마이크로데이터(micro data)라 부르는데, 이러한 마이크로데이터에 대한 기밀보호(confidentiality protection)가 국내적인 문제는 물론, 인터넷을 통한 데이터 전파가 활성화되고 있는 오늘날은 국제적인 문제로 되고 있다.

지난 10여년에 걸쳐 정부와 관련을 맺고 있는 대학이나 연구기관이 늘고 있다. 지방자치 단체가 설립하여 운영하는 연구소도 질적 양적 양면으로 큰 성장을 이루고 있다. 또한 민간부문에서도 세계시장을 선도하는 기업들이 늘어나고, 이러한 기업들은 자체 설립 연구소를 통해 세계 시장의 흐름을 판단하고 미래전략을 세우고 있다. 그러나 이상의 대학, 연구기관, 지방자치단체 연구소, 기업 연구소가 마이크로데이터를 제공받지 못하고 집계수준의 정보만 제공받고서는 세밀한 연구를 할 수 없게 된다. 예로, 정부에서 생활보호대상자를 선정하는 기준을 마련하고자 하는 경우 마이크로데이터인 각 가구의 가구구성원별 소득을 먼저 알아야 하고 다음으로 이를 가구별로 집계하여야 한다. 마이크로데이터를 공급받지 못하고 대신 집계데이터만 공급받아 연구한다면 올바른 연구를 할 수 없게 되는 것이다. 이러한 연유로 최근 국내에서도 마이크로데이터의 수요가 급증하고 있으며, 또한 세계화라는 시대적 조류 때문에 나라 간에서도 마이크로데이터의 상호제공 필요성이 늘어나고 있는 것이다[3].

반면 기밀보호 면으로 보면 다른 문제가 발생하게 된다. 통계 기밀보호에 관한 UN의 6번째 원칙을 살펴보면 다음과 같다.

"자연인이든 법인을 대상으로 하든 간에 통계기관이 수집한 개별 데이터 즉 마이크로데이터는 철저히 기밀보호가 되어야 하며, 전적으로 통계목적으로만 쓰여야 한다."

이러한 기밀보호 원칙을 지키자면 마이크로데이터의 활용에 제약이 생기고, 결과적으로 집계 데이터가 더 많이 쓰이게 된다. 전술한 바와 같이 집계 데이터만 사용하면 세밀한 연구를 할 수 없게 된다. 마이크로데이터를 사용하여 통계정보를 도출할수록 통계정보는 왜곡되지 않으며, 반대로 집계 데이터를 사용하여 통계정보를 도출할수록 통계정보는 왜곡되게 된다. 따라서 될 수 있는 대로 마이크로데이터에 개념적으로 가까운 집계 데이터를 제공하고 사용하게 하면 UN의 6번째 원

칙을 위배하지 않으면서도 정부정책 입안과 시행 및 대국민 홍보의 정부 목적을 효과적으로 달성할 수 있으리라 본다.

마이크로데이터를 저장하고 있는 한 개의 핵심 큐보이드(core cuboid)와 다차원 수준의 조합별 집계 데이터를 저장하고 있는 여러 개의 집계 큐보이드로 구성되는 것이 OLAP(OnLine Analytical Processing) 데이터 큐브(data cube)[4]인데, 이 데이터 큐브의 일부 즉 몇 개의 집계 큐보이드를 공개하는 식으로 해서 허가 받은 연구자에게든 일반 대중에게든 공식통계를 공개하는 것이 오늘날 가장 발전된 통계 공개방식 중의 하나이다. 그런데, 집계 큐보이드를 아무런 기준 없이 공개하다 보면 이렇게 공개된 집계 큐보이드만 분석하드라도 데이터 큐브의 핵심 큐보이드에서의 데이터와 같은 중요 데이터를 추론해 낼 수 있게 된다는 심각한 기밀보호 상의 문제가 생기게 된다[5]. 데이터 큐브의 중요 데이터가 누설되지 않도록 통제하는 것을 추론통제(inference control)[6]라 부르는데, 본 연구에서는 특허출원 분야의 공식통계를 사례로 삼아 연구함으로써 정부정책 입안과 시행 및 대국민 홍보의 정부 목적을 달성할 수 있는 공식통계의 추론통제 전략을 제안하는 것을 목적으로 하고 있다.

II. 관련 연구

자연인이나 법인 그룹을 대상으로 한 빈도, 평균값과 같은 여러 통계 값을 공표함과 동시에 이들 통계 값의 집계출처가 되는 마이크로데이터의 기밀을 보호하고자 하는 목적으로 1980년대부터 크게 주목받기 시작한 것이 통계 데이터베이스(statistical database: 이하 SDB로 기재)이다. 데이터에 내재되어 있는 특이패턴을 찾고자 데이터 분석을 할 때에 보통 다차원적인 데이터 집계를 하는데, 이때에 표준 SQL(Structured Query Language) 쿼리를 사용해도 좋지만 쿼리가 아주 복잡해진다는 단점이 생기게 된다. 쿼리가 복잡해지면 SDB의 근본 구성데이터 즉 베이스 데이터(base data)를 여러 번 참조해야 되고 결과적으로 쿼리의 성능이 저하되게 된다. 오늘날은 우리가 사용하는 의사결정용 쿼리가 복잡한 게 대다수이기 때문에 표준 SQL 쿼리를 계속 사용할 수 없는 실정이다. 따라서 이를 대신할 새로운 집계용 연산자인 OLAP 데이터 큐브가 1990년대에 도입되기 시작했으며[6], 오늘날 이 데이터 큐브를 네덜란드의 중앙통계국 CBS(Centraal Bureau voor de Statistiek)를 위시해 여러 기관이 통계정보를 공표하는 주력 매개물로 삼고 있는 것이다[7].

추론통제기법 중 어떤 것이 좋은 가는 보안성, 정보손실 그리고 비용이라는 3가지 인자를 평가함으로써 결정하게 된다. 여기서 보안성이란 해당 추론통제기법을 실행시켰을 경우 유추되는 중요 데이터의 양에 따라 그 높고 낮음이 결정되고, 정보손실이란 중요하지 않는 데이터까지도 얼마나 많이 공표할 수 없도록 억제시키느냐에 따라 그 많고 적음이 결정되며, 비용은 쿼리 처리에 따라 결정된다. 유추되는 즉 기밀보호가 되지 못하는 중요 데이터의 양이 적을수록 보안성이 높아지고, 중요하지 않는 데이터를 많이 공표할수록 정보손실이 적어져서 비용까지 적게 들면 아주 우수한 추론통제기법이 되는 것이다. 그러나 보안성을 높이면 정보손실이 많아지고 고비용이 되기 때문에 이 3가지 인자 간에 적정균형을 맞춰주는 추론통제기법을 선택할 필요가 있다고 본다[8,9].

오늘날 OLAP 데이터 큐브에서의 추론통제기법이 여럿 개발되고 있으나 결과물 제약통제(output restriction controls) 기법[6]과 같은 SDB에 적용한 기법을 그대로 OLAP 데이터 큐브에 적용시키는 게 대부분이다. 그러나 OLAP 데이터 큐브의 특성을 고려하지 않고 SDB에서의 추론통제방식을 데이터 큐브에 직접 적용시키면 데이터 큐브는 올바른 정보를 싣지 못한 채 다른 말로 하면 왜곡된 채 추론통제를 받게 되는 모순에 빠질 수도 있게 된다. 정부의 연구개발예산이나 각종 평가 및 심사 등에 관한 국가통계는 주로 평균값과 횟수로 공표되고 있는 점을 감안하면 집계함수 중 평균 집계함수 AVG와 셈 집계함수 COUNT가 아주 중요하다고 볼 수 있는데, 전술한 결과물 제약통제기법을 데이터 큐브에 적용시킬 경우에는 합 집계함수 SUM을 사용할 때에만 정당성을 얻을 수 있고 AVG와 COUNT를 사용할 때에는 정당성을 얻지 못하게 된다. 왜냐하면 SDB에서와는 달리 데이터 큐브에서는 태생적으로 결측값(missing value)이 생길 수밖에 없고 이 때문에 COUNT에 오류가 생기고 더불어 AVG에도 오류가 생겨서 결과적으로 데이터 큐브가 왜곡되기 때문이다[10,11].

Wang 등[6]은 집계함수의 사용에 제약이 있는 결과물 제약통제기법을 대체할 새로운 기법을 보고하고 있다. 이 기법은 분배적 집계함수를 사용한다면 집계함수가 어떤 종류의 것이든 전혀 신경 쓰지 않아도 되는 현실적인 기법이라고 하고 있다. Palpanas[12]와 Gray[13] 그리고 Wang은 셈 집계함수 COUNT를 분배적 함수로 보고 있다. 특히 Wang은 데이터 큐브를 작성하는데 관련된 집계함수 중 대수적 함수인 평균 집계함수 AVG는 매개함수 개념을 활용하면 데이터 큐브 작성에 직접 사용할 수 있다고 하였다. AVG 자체는 분배적 함수가 아니지만, (SUM, COUNT) 매개함수가 분배적 함수이기 때문에 AVG를 (SUM, COUNT)를 통해 구하면

된다고 한 것이다[14,15].

그러나 저자들은[16] 최근, 집계함수의 종류에 구애받지 않고 추론통제를 할 수 있다는 Wang의 추론통제기법에 오류가 있음을 확인하고 이를 바로잡는 새로운 체제를 구축하였다. 구체적으로 현 OLAP 체제에서는 COUNT의 기능이 올바로 발휘되지 못한다는 점과 COUNT를 활용해야만 계산할 수 있는 AVG의 기능도 올바로 발휘되지 못한다는 점을 확인하고, 이를 해결할 체제를 구축한 것이다. 올바른 정보를 갖고 있는 즉 왜곡되지 않은 데이터 큐브를 작성하는 것이 먼저이고, 이 큐브로부터 정보가 누설되지 않도록 통제하는 것은 다음이다. 다시 말하면 SDB에는 올바른 정보가 입주해 있기에 SDB로부터 정보가 누설되는 것을 막는 추론통제를 논하는 것이 타당하지만, OLAP 데이터 큐브에는 애초부터 올바른 정보가 입주할 수 없는 경우가 생길수도 있기 때문에 이렇게 왜곡된 데이터 큐브로부터의 추론통제를 논하는 것은 무의미하다는 뜻이다[10]. 저자들은 추론통제에 앞서 먼저 올바른 평균값 정보를 지니게 되는 데이터 큐브를 작성하고, 다음으로 이 큐브에서 정보가 누설되는 것을 막는 보안성에 있어서 신축성이 있으며 정보손실과 비용이 적게 드는 새로운 추론통제기법을 개발한 것이다. 본 연구에서는 공식통계의 추론통제 전략을 세우는 한 가지 방법을 이 추론통제기법을 활용하여 나타내고자 한다.

III. 정부의 특허경비지원사업

3.1 연구개발사업에 관한 정보공개

지난 날 정부의 연구개발사업은 연구개발에 대한 체계적인 관리가 없이 연구자에게 모든 활동을 일임하고, 막연하게 성과를 기대하는 수준이었다. 즉, 정부는 막대한 예산을 투입하고, 연구자의 자유로운 활동을 보장하는 것이 우수한 기술을 개발할 수 있는 최고의 방법이라고 생각했었다. 그러나 이는 한가지 기술만 사용하여 제품을 생산하고 판매하는 시장에서나 성공 가능했던 과거의 연구개발 형태였다.

최근에는 개인의 취향이 다양해지고 제품에 대한 요구사항이 많아지면서, 하나의 제품을 출시하기 위해선 여러 기술을 사용해야만 하는 상황으로 급변하고 있다. 이러한 급변상황에 부응하기 위해서는 예산뿐만 아니라 연구인력, 연구기자재, 연구성과 등 연구개발 자원 전반에 걸쳐 관리해야만 하는데, 이를 위해 정부는 국가과학기술지식정보서비스(NTIS, National Science & Technology Information Service)를 운영하고 있다. 연구개발의 기획에서 활용에 이르기까지 전 주기에 걸쳐 연구개발의 효용성을 높이기 위해 국가 연구개발 사업정보를 구축하고 제공하는 서비스가 NTIS이다[17].

정부 연구개발 사업정보의 공개 수준을 알아보기 위해 연구개발사업의 추진체계를 살펴보고자 한다. 정부 연구개발사업은 매년 사업의 통합공고로써 시작되는데, 이 정보는 대중매체, 홍보자료 및 정부 연구개발투자 부처 합동설명회(이하 합동설명회로 기재)를 통하여 공표된다. 합동설명회에서는 각 부처의 사업별 정책방향 및 예산규모 등의 포괄적인 정보를 제공하고, 다음으로 연구관리 전문기관(이하 전문기관으로 기재)의 세부사업 공고에 의하여 사업범위와 예산이 공고되고 있으며, 이 공고에 따라 기업들의 사업신청이 이루어지고 있다.

전문기관에서는 각 사업공고를 통하여 사업의 목적, 내용, 예산, 지원대상 및 조건 등의 비교적 자세한 정보를 제공하지만, 이는 기업이 연구개발사업에 참여할 수 있는 최소한의 자격을 알리는 정보에 불과하며, 선정되는 기업이 지녀야 할 능력을 지정하는 정보는 아니다. 기업은 각 부처가 공개하는 통계 및 홍보자료를 통하여 해당 사업의 평균경쟁률, 평균지원액, 선정기관 수 등의 정보를 추가로 획득할 수 있으나 이 모든 공개된 정보를 종합하드라도, 기업은 어느 부분에서 능력이 부족하여 선정되지 못했는지 정확하게 판단할 수 없는 것이 현 실정이다.

그러나 현재 연구개발사업에 대한 정보공개는 합동설명회, 전문기관의 사업공고, 각 부처의 통계 및 홍보자료의 공개 같은 것에 크게 의존하고 있으며, 각 기업을 평가한 구체적 정보는 공개하지 않는 것을 원칙으로 하고 있어, 기업의 부족한 면을 발굴 ·개발할 수 있는 동기부여를 원천봉쇄하고 있는 상황이다. 이와 같이 기업에 구체적 정보를 제공하는 것이 필요함에도 불구하고, NTIS는 개인정보보호와 기업보안을 이유로 구체적 정보를 제공하지 않고 있다. 통계 기밀보호에 관한 원칙을 지키는 선에서 최대한 많은 정보를 공개하는 것이 정부나 기업 모두에 좋다. 정부로서는 정부정책 시행의 투명성을 높여서 좋고, 기업은 자신의 부족한 점을 파악하여 대처할 수 있는 동기부여를 받아 좋다.

3.2 특허경비지원사업

본 연구에서는 정부의 특허경비지원사업(The Patent Application Expense Support Project, 이하 특허사업으로 기재)을 사례로 삼고자 한다. 특허사업은 출연연구소, 국공립연구소, 대학 등 공공기관이 적시에 특허출원과 등록을 진행하도록 특허출원 경비를 지원하고 관련 정보를 공개함으로써 기술이전 및 사업화 활동을 촉진시키는

사업이다.

정부 연구개발사업은 연구비에서 특허출원 및 등록비를 사용할 수 있으나, 해외출원은 1개국당 약 천만원 정도가 소요되어 경비를 충당하기 어려운 상황이며, 연구비에서 특허경비를 과다 편성할 경우 실제 연구비가 줄어들어 연구개발 활동에 지장을 초래할 수 있다. 또한 연구비는 연구기간 내에 지출과 정산을 해야 하고, 특허는 연구과제의 종료 후 연구성과로 나타나는 특성상, 특허경비를 연구비로 책정하는 것은 실제로 불합리한 실정이다. 특히, 공공기관의 연구개발 결과는 대부분 원천기술에 해당되어 장기적인 연구가 필요하며, 이러한 원천기술은 국내뿐만 아니라 해외출원이 필수적이고, 최근 국가의 과학기술 경쟁력 확보, 기술유출 방지 및 미래성장 동력의 창출이라는 정책방향을 감안할 때에 매우 중요한 사업이라 할 수 있다. 특허사업은 2000년부터 2008년까지 매년 2회 추진되었으며, 이 기간 동안 총 89개의 기관(출연 및 국공립연구소 20개, 대학 69개)이 전세계 23개의 나라에 특허를 출원・등록하였으며, 그 규모는 총 219억원에 달하는 사업이다.

최근의 세계경제는 자유무역협정(FTA, Free Trade Agreement), 즉 특정국가간에 배타적인 무역 특혜를 서로 부여하는 지역경제 통합 형태로서 재편되고 있고[18], 지역경제권은 지정학적 분류에 따라 유럽권, 아시아경제권, 아메리카경제권 또는 경제교역에 따라 동북아시아경제권, 환태평양경제권 등으로 구분되어진다. 특히 FTA로 대표되는 지역주의는 타 경제권에 배타적이어서 이 경제권내에 통용될 수 있는 제품의 생산거점을 확보하고, 사업화 가능성이 높은 특허를 경제규모가 큰 나라에 집중 출원하는 전략 등은 매우 중요하다 할 수 있다. 그러나 특허사업에 대한 공개정보가 출원건수, 등록 건수 같은 것으로 제한되어 있어 기업이나 공공기관이 출원전략을 수립하는 데에 어려움을 겪고 있다.

특허출원 전략을 수립하는 데에는 구체적 정보, 즉 전체 특허출원경비, 평균 특허출원경비, 출원국가, 기관 등에 대한 상세한 통계정보가 필요함에도 불구하고 공공기관의 정보보호와 정부의 사업보안을 이유로 구체적 정보는 제공되지 않고 있다. 본 연구에서는 특허사업의 경제권역별 세부전략 수립을 위하여 지정학적 분류에 따른 대륙별 구분을 적용하였으며, 기관 및 나라의 경쟁력 비교를 위하여 평균 특허출원경비를 분석하고 있다.

3.2.1 특허사업

본 사업이 추진되었던 2000년부터 2008년까지의 9년 동안 각 기관이 수혜한 특허출원경비를 분석하고자 1개의 사실 테이블과 3개의 차원 테이블을 채택하여 다음과 같은 스타 스키마를 구축하였다.

차원 테이블로는 시간(TIME) 테이블, 국가(COUNTRY) 테이블, 기관(ORGANIZATION) 테이블의 3개가 있다. TIME 차원 테이블은 정부사업을 추진한 기간을 나타내며, all, 연(Year), 반기(Half_Year)의 3개 속성으로 구성하였다. 각 속성의 구성원은 all은 1개, 연은 9개, 반기는 각 년마다 상반기 및 하반기 2개로써 18개가 존재하며, 기본키를 구성하는 Time_ID 18개를 부여하여 각 행을 식별하였다. 이렇게 구성된 TIME 차원 테이블은 표 1과 같다.

표 1. TIME 차원 테이블
Table 1. TIME Dimension Table

Time_ID	Half_Year	Year	all
1	1	2000	ALL
2	2	2000	ALL
3	1	2001	ALL
4	2	2001	ALL
5	1	2002	ALL
⋮	⋮	⋮	⋮
18	2	2008	ALL

COUNTRY 차원 테이블은 각 기관이 특허 출원한 나라를 나타내며, all(세계), 대륙(Continent), 나라(Nation)의 3개 속성으로 구성하였다. all 속성에는 1개, Continent 속성에는 Asia, Europe, Oceania, Central America, North America, PCT Continent 라는 6개의 구성원이 있다. 또한 Nation의 구성원으로서 Asia의 구성원은 China, India, Japan, Korea, Philippine, Singapore, Taiwan이 있으며, Europe의 구성원은 England, EPO(유럽특허청, European Patent Office), Finland, France, Germany, Italy, Netherlands, Russia, Spain, Switzerland가 있으며, Oceania의 구성원은 Australia, NewZealand가 있고, Central America의 구성원은 Mexico가 있고, North America의 구성원은 Canada, USA가 있고, PCT Continent의 구성원으로는 PCT(특허협력조약, Patent Cooperation Treaty) 등 총 23개가 있으며, 기본키로 사용할 Country_ID 23개를 부여하여 각 행을 식별하도록 하였다. COUNTRY 차원 테이블은 표 2와 같다.

표 2. COUNTRY 차원 테이블
Table 2. COUNTRY Dimension Table

Country_ID	Nation	Continent	all
1	China	Asia	ALL
2	India	Asia	ALL
3	Japan	Asia	ALL
4	Korea	Asia	ALL
5	Philippine	Asia	ALL
⋮	⋮	⋮	⋮
23	PCT	PCTContinent	ALL

ORGANIZATION 차원 테이블은 특허사업으로 9년 동안 한번이라도 지원받은 기관으로서, 기관 위치에 따른 ORGANIZATION(Ⅰ) 차원과 기관 성격에 따른 ORGANIZATION(Ⅱ) 차원으로 구분하여 구성하였다. ORGANIZATION(Ⅰ) 차원은 3개의 속성으로 구성되며, all 속성에는 1개, Category 속성에는 11개, Name 속성에는 89개의 구성원이 존재한다. all 속성은 모든 Category를 통합한 것이고, Category 속성은 기관 위치에 따라 Chungcheong, Daejeon, Incheon, Jeju, Jeolla, Kangwon, Kwangju, Kyeongsang, Kyunggi, Pusan, Seoul로 11개의 구성원이 있다. 또한 Name 속성은 지원받은 기관으로 Chungcheong에는 4개 구성원이, Daejeon에는 17개 구성원이, Kwangju에는 4개 구성원이, Kyunggi에는 9개 구성원이, Seoul에는 24개 구성원이, Incheon에는 4개 구성원이, Jeju에는 1개 구성원이, Jeolla에는 8개 구성원이, Kwangwon에는 4개 구성원이, Kyeongsang에는 8개 구성원이, Pusan에는 6개 구성원이 있다. 여기에 Organization_ID 89개를 부여하여 각 행을 식별하여 기본키로 사용한다. ORGANIZATION(Ⅰ) 차원 테이블은 표 3의 (I)과 같다. 그리고 각 기관의 익명성을 보장하기 위하여 기관명은 Category명_Class명_일련번호로 수정하여 기재하였다. 예로 충정지역에 소재하는 연구소의 경우 CC_I01로, 대학의 경우 CC_U01로 표기하였다.

ORGANIZATION(Ⅱ) 차원 테이블은 ORGANIZATION(Ⅰ) 차원의 Category 속성 대신에 기관 성격을 나타내는 Class 속성으로 대체한 차원이다. all 속성에는 1개, Class 속성에는 2개, Name 속성에는 89개의 구성원이 존재한다. Class 속성은 기관 성격에 따라 구분한 것으로 Institute에는 20개 구성원이, University에는 69개 구성원이 있다. 여기에 Organization_ID 89개를 부여하여 각 행을 식별하여 기본키로 사용한다. ORGANIZATION(Ⅱ) 차원 테이블은 표 3의 (II)와 같다.

표 3. ORGANIZATION 차원 테이블
Table 3. ORGANIZATION Dimension Table

Organization_ID	Name	Category	all
1	CC_I01	Chungcheong	ALL
2	DJ_I01	Daejeon	ALL
3	DJ_I02	Daejeon	ALL
4	DJ_U01	Daejeon	ALL
5	DJ_I03	Daejeon	ALL
6	DJ_I04	Daejeon	ALL
7	DJ_I05	Daejeon	ALL
⋮	⋮	⋮	⋮
89	SU_U21	Seoul	ALL

(Ⅰ)

Organization_ID	Name	Class	all
1	CC_I01	Institute	ALL
2	DJ_I01	Institute	ALL
3	DJ_I02	Institute	ALL
4	DJ_U01	University	ALL
5	DJ_I03	Institute	ALL
6	DJ_I04	Institute	ALL
7	DJ_I05	Institute	ALL
⋮	⋮	⋮	⋮
89	SU_U21	University	ALL

(Ⅱ)

사실 테이블은 TIME 차원, COUNTRY 차원, ORGANIZATION 차원의 기본키로 구성된 복합키 부분과 측정값으로 나타내는데, 특허출원경비를 나타내는 측정값은 ApplicationExpense이다. 표 4는 TIME 차원은 2000년 상반기, ORGANIZATION 차원은 Kwangju 지역의 사실테이블을 나타낸 것이다.

표 4. 특허 사실 테이블
Table. 4. The Patent Fact Table

	Category / Name / Nation	Kwangju			
		KJ_I01	KJ_U01	KJ_U02	KJ_U03
Asia	China	-	3,715,252	-	-
	India	-	-	-	-
	Japan	-	33,420,575	-	-

	Korea	-	34,612,700	-	-
	Philippine	-	-	-	-
	Singapore	-	-	-	-
	Taiwan	-	-	-	-
Europe	England	-	6,285,827	-	-
	EPO	-	17,744,393	-	-
	Finland	-	-	-	-
	France	-	3,980,905	-	-
	Germany	-	3,320,482	-	-
	Italy	-	-	-	-
	Netherlands	-	-	-	-
	Russia	-	-	-	-
	Spain	-	-	-	-
	Switzerland	-	-	-	-
Ocenia	Australia	-	-	-	-
	NewZealand	-	-	-	-
Central America	Mexico	-	-	-	-
North America	Canada	-	-	-	-
	USA	-	41,794,191	-	-
PCT Continent	PCT	-	-	-	-

3.2.2 데이터 큐브

3.2.1의 특허사업 사례를 데이터 큐브로 나타내면 그림 2과 같은데, 그림 2의 데이터 큐브에는 TIME, COUNTRY, ORGANIZATION이라는 3개의 차원(dimension)이 있다. TIME 차원은 Half_Year, Year, all 이라는 3개의 속성(attribute)으로 구성되고, COUNTRY 차원은 Nation, Continent, all 이라는 3개의 속성으로 구성되며, ORGANIZATION(Ⅰ) 차원은 Name, Category, all 이라는 3개의 속성으로, ORGANIZATION(Ⅱ) 차원은 Name, Class, all 이라는 3개의 속성으로 구성된다. 이와 같은 각 차원의 속성을 나타내면 그림 1과 같다.

본 연구에서는 그림 1(I)에서의 TIME, COUNTRY, ORGANIZATION(I) 차원을 갖고 데이터 큐브(I)을 작성하고, 그림 1(II)에서의 TIME, COUNTRY, ORGANIZATION(II) 차원을 갖고 데이터 큐브(II)를 작성하고자 하는 것이다.

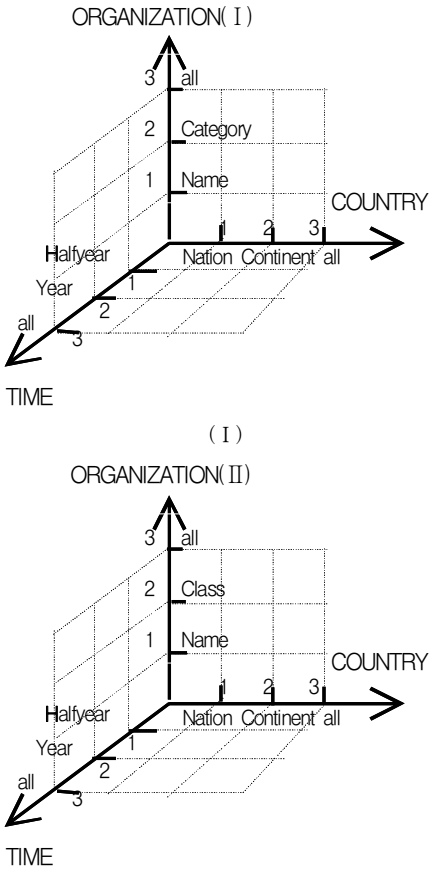


그림 1. 각 차원의 속성
Fig. 1. The Attributes of Each Dimension

그림 1의 (I), (II) 어디에서 보든 간에 이들 3개 차원의 각 속성간의 교차점을 구해보면 27개(3×3×3)가 되는데, 이들 27개 각각의 교차점에서 하나씩의 큐보이드 〈TIME, COUNTRY, ORGANIZATION〉가 생성된다. 각 차원의 속성 이름 대신에 해당 숫자를 사용하여 종속관계를 보여주는 총 27개의 큐보이드로 구성된 데이터 큐브를 나타낸 것이 그림 2이다. 그림 2의 사례 데이터 큐브는 데이터 큐브(I)이 될 수도 있고 데이터 큐브(II)도 될 수 있음을 확인하기 바란다.

그림 2의 사례 데이터 큐브를 데이터 큐브(I)이라 볼 때에 일례로 큐보이드 〈2, 2, 2〉는 TIME 차원의 Year 속성, COUNTRY 차원의 Continent 속성, 그리고 ORGANIZATION(I) 차원의 Category 속성이 교차하는 점에서 생성된다. 또한 이

큐보이드에는 594개(Year 개수 9 × Continent 개수 6 × Category 개수 11)의 셀(cell)이 마련되는데 여기에 지원금 집계 데이터가 수록되는 것이다. 또한 핵심 큐보이드(core cuboid)는 〈1,1,1〉인데 이를 문자로 표시하면 〈Half_Year, Nation, Name〉이다. 이 핵심 큐보이드의 각 셀에 측정값 속성인 ApplicationExpense를 입력할 때에 사용하는 테이블을 베이스 테이블이라 하는데, 본 사례에서의 베이스 테이블(base table) 스키마(schema)는 (Half_Year, Nation, Name, ApplicationExpense)이다.

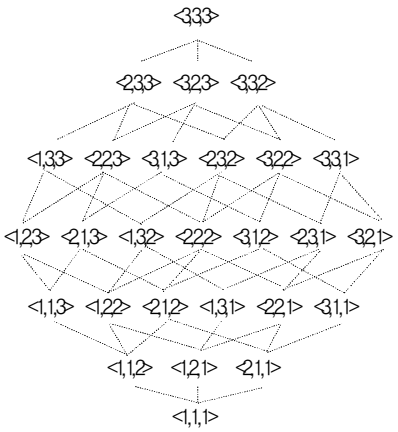


그림 2 사례 데이터 큐브
Fig. 2. The Case Data Cube

Ⅳ. 공식통계의 추론통제 모델

저자들은 평균 집계함수를 사용하여 데이터 큐브를 작성할 때에 있어서 먼저 올바른 평균값 정보를 지니게 되는 데이터 큐브를 작성하고, 다음으로 이 큐브에서 정보가 누설되는 것을 막는 보안성에 있어서 신축성이 있으며 정보손실과 비용이 적게 드는 새로운 추론통제기법을 개발한 바 있다[16].

본 연구에서는 3.2.1의 특허사업을 사례로 삼아, 이 새로운 추론통제기법을 활용하여 공식통계의 추론통제 전략을 세우는 한가지 방법을 논하고자 한다.

경제권역별로 특허출원 세부전략을 수립하고 각 기관 및 우리나라의 경쟁력을 알아보기 위하여 각 기관이 지원받은 특허출원경비를 OLAP 베이스 테이블의 측정값 속성으로 설정하고, 이 OLAP 베이스 테이블에 정확한 통계값을 입주시켜서 추론통제를 하는 새로운 통제기법을 단계별로 설명하면 다음과 같다.

4.1 추론통제 모델

〔1단계〕 공백 값 상태의 사실테이블 활용(표 4)

표 4는 특허사업에 관련된 마이크로데이터를 기입한 사실테이블이다. 표 4에서와 같이 공백값 상태의 사실 테이블에 기초하여 데이터 큐브를 작성하고 이로부터 평균값 정보를 구해보면 오류가 생긴다. 이러한 오류를 없애자면 그림 4과 같은 결측값 처리를 한 베이스 데이터에 기초하여야 한다고 밝힌 바 있다.

그러나 OLAP에서 그림 4과 같은 베이스 데이터를 입력하는 것은 현실적으로 불가능하며, 경제적이지 못하기 때문에 우선 현실적인 공백값 상태의 사실테이블을 데이터 큐브 작성의 기초로 삼아 출발하고, 〔2단계〕에서 결측값 처리를 한 사실테이블로 전환시킨다.

〔2단계〕 베이스 테이블의 변환(표 4 → 그림 4)

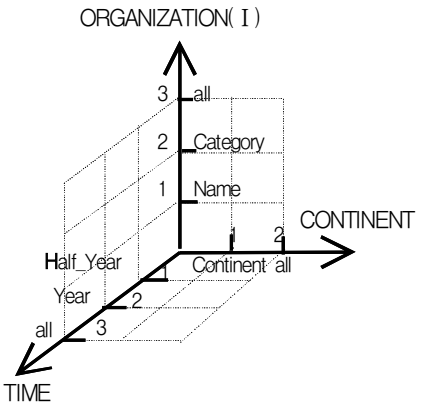
본 사례 연구에서는 우리나라 각 기관이 특허출원을 얼마나 왕성하게 하느냐를 척도로 삼아 각 기관의 국제경쟁력을 평가하고자 한다. 예를 들어 표 4에서의 Kwangju 지역에 있는 기관 KJ_U01의 특허출원경비를 살펴보면 China에 3,715,252원, Japan에 33,420,575원, Korea에 34,612,700원의 경비를 들인 것으로 되어 있다. 현 OLAP 체제하에서 KJ_U01 기관의 평균 특허출원경비를 계산하면 총 합계 71,748,527원을 3으로 나눈 23,916,176원으로 계산 되는데, 이는 잘못되었다고 본다. 왜냐하면, Asia 대륙에 속하는 India, Philippine, Singapore, Taiwan의 네 나라에는 전혀 출원을 하지 않았다는 사실을 감안하지 않았기 때문이다. 진정한 의미에서의 KJ_U01 기관의 국제경쟁력을 평가하자면 출원한 3 나라만 대상으로 해서는 안되고 Asia 대륙에 속해있는 7 나라를 모두 대상으로 해야 한다고 본다. 즉 KJ_U01 기관이 7개국에 출원을 했다고 했을 때에 1개국 당 평균적으로 얼마만큼의 출원을 했는가를 알아보아야 하며, 이를 위해서는 총 합계 71,748,527원을 3이 아닌 7로 나누어야 한다는 뜻이다. 따라서 KJ_U01 기관의 평균 특허출원경비는 10,249,790원으로 산출하게 된다. 본 단계에서는 이와 같은 올바른 평균 특허출원경비를 산출하기 위해 다음과 같은 세부단계를 거침으로써 베이스 테이블 자체를 표 4에서 그림 4로 변환시키고자 하는 것이다.

〔2-1 단계〕 특허출원을 하지 않는 대륙에 있어서의 기관의 평균 특허출원경비는 0으로 기입한다.

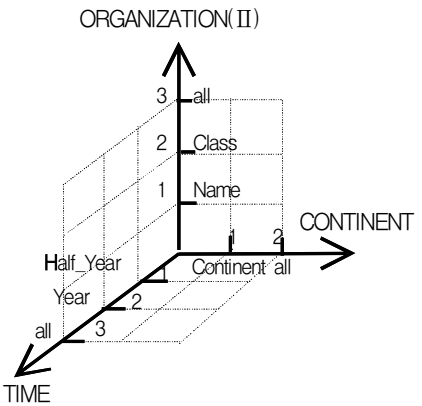
〔2-2 단계〕 특허출원을 한 대륙에 있어서의 기관의 평균 특허출원경비는 총 특허출원경비를 각 대륙에 속하는 나라의 수인 고정 COUNT 값으로 나누어 기입한다. 대륙별 고정 COUNT 값은 Asia는 7, Central America는 1, Europe

은 10, North America는 2, Oceania는 2, PCT Continent는 1이 된다.

대륙별 각 기관의 평균 특허출원경비를 구하자면 COUNTRY 차원에서의 Nation 속성은 분석대상에 넣을 필요가 없다. 즉, 그림 1에서의 COUNTRY 차원의 Nation 속성을 제거시킨 그림 3과 같은 새로운 차원을 구성해야 한다. 또한 그림 4에서 보다시피 새로운 베이스 테이블 스키마는 〈Half_Year, Continent, Name, AvgApplicationExpense〉로 된다. 이때의 행의 개수는 9,612(TIME 18개 × CONTINENT 6개 × ORGANIZATION 89개)가 된다.



(I)



(II)

그림 3. 변경된 차원의 속성
Fig. 3. The Attributes of Modified Dimensions

TimeID	ContinentID	OrganizationID	AvgAppExp
1	1	1	0
1	1	2	0
1	1	3	0
1	1	4	69275451.4285714
1	1	5	0
1	1	6	0
1	1	7	0
1	1	8	0
1	1	9	0
1	1	10	2538199
1	1	11	110871.428571429
1	1	12	815714.285714286
1	1	13	0
1	1	14	7614368.28571429
1	1	15	541371.428571429
1	1	16	10249789.5714286
1	1	17	0
1	1	18	0
1	1	19	0
1	1	20	171071.428571429
1	1	21	0
1	1	22	37437950.9142857
1	1	23	0

TimeID	ContinentID	OrganizationID	AvgAppExp
1	1	24	707357.142857143
1	1	25	0
1	1	26	487286.428571429
1	1	27	0
1	1	28	0
1	1	29	0
1	1	30	366471.428571429
1	1	31	0
1	1	32	0
1	1	33	0
1	1	34	472042.857142857
1	1	35	0
1	1	36	0
1	1	37	0
1	1	38	0
1	1	39	0
1	1	40	0
1	1	41	0
1	1	42	0
1	1	43	0
1	1	44	0
1	1	45	0
1	1	46	0

그림 4. 새로운 특허 사실 테이블
Fig. 4. The Modified Patent Fact Table

〔3단계〕 새로운 데이터 큐브 구축(그림 5)

그림 3에 표시한 각 차원과 그림 4의 사실 테이블을 갖고 구성한 데이터 큐브의 격자구조를 살펴보면 그림 5와 같다〔19〕. 여기서 핵심 큐보이드는 〈1,1,1〉 즉 〈Half_Year, Continent, Name〉이다. 여기서 그림 5를 데이터 큐브(I)이라 보고 2000년도 1반기에 해당되는 큐보이드 〈1,2,2〉와 이의 선조 큐보이드 〈1,1,2〉, 〈1,2,1〉, 〈1,1,1〉을 나타낸 것이 그림 6이다.

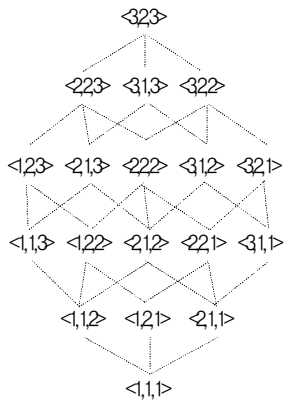


그림 5. 새로운 특허 큐브의 격자구조
Fig. 5. The Lattice Structure of the New Patent Data Cube

Year	Halfyear	Category				
2000	1	Kwangju				
	Name					
	KJ_I01	KJ_U01	KJ_U02	KJ_U03	총합계	
Continent	AvgAppExp	AvgAppExp	AvgAppExp	AvgAppExp	AvgAppExp	
Asia	0	10,249,790	0	0	10,249,790	
CentralAmerica	0	0	0	0	0	
Europe	0	3,133,161	0	0	3,133,161	
NorthAmerica	0	20,897,096	0	0	20,897,096	
Oceania	0	0	0	0	0	
PCTContinent	0	0	0	0	0	
총합계	0	34,280,046	0	0	34,280,046	

그림 6. 큐보이드 〈1,2,2〉와 선조 큐보이드
Fig. 6. A Cuboid 〈1,2,2〉 and Its Ascendant Cuboids

〔4 단계〕 중요 정보의 지정(Below(〈2,1,1〉))

3.2.1의 특허사업 사례에서 대륙별 각 기관의 평균 특허출원경비, 즉 연도별, 대륙별 기관의 특허출원경비 정보보다 더 개괄적인 정보는 공표해도 되지만 그렇지 않은 경우는 절대적으로 보호해야 한다는 본 연구에서 방침을 고려해보자. 그림 7에서 볼 것 같으면 연도별, 대륙별, 기관별 특허출원경비 정보를 가지고 있는 것은 큐보이드 〈2,1,1〉이며, 이 큐보이드보다 더욱 상세한 정보(예를 들면, 반기별, 대륙별 기관의 평균 특허출원경비를 포함하고 있는 큐보이드)를 가지고 있는 큐보이드는 보호해야 하는데, 이 큐보이드를 선조 큐보이드라고 부르며, 큐보이드 〈1,1,1〉이 해당된다. 따라서 보호해야 할 큐보이드는 〈2,1,1〉, 〈1,1,1〉 이며, 이는 Below(〈2,1,1〉)로 나타내며, 타깃이라 부른다.

Below(〈2,1,1,〉) = {〈2,1,1〉, 〈1,1,1,〉}

즉, 타깃내의 큐보이드는 보호해야 할 큐보이드 〈2,1,1〉보다 더 자세한 정보를 가지고 있기 때문에, 타깃내의 큐보이드가 공표되었을 경우에는 중요 정보의 보호가 불가능하게 되는 것이다.

〔5단계〕 새로운 데이터 큐브에서의 추론

그림 7에서 Below(〈2,1,1〉)은 보호 큐보이드 집합으로서, 실선 ① 아래에 있는 큐보이드가 이에 해당하며, 타깃이 된다. 이 타깃내의 큐보이드 만을 보호하였을 경우 보호 큐보이드의 자손으로부터 추론이 가능하게 된다. 즉 실선 ① 위에 있는 큐보이드 집합을 소스(source, 이하 S로 표기)라 부르는데, 이 S의 셀 데이터를 잘 활용하면 보호 큐보이드 집합의 셀 데이터를 추론해낼 수 있다.

타깃의 큐보이드 〈2,1,1〉과 이의 직계후손인 큐보이드로서 S에 속해있는 3개의 큐보이드 〈2,1,2〉, 〈2,2,1〉, 〈3,1,1〉에 대해 알아보자. 3개의 큐보이드 〈2,1,2〉, 〈2,2,1〉, 〈3,1,1〉 각각은 서로 간의 선조-후손의 관계를 규정할 수 없기 때문에, 이들 각자 큐보이드의 셀은 선조 큐보이드 〈2,1,1〉의 추론에 기여를 할 수 없다. 그러나 이들 3개의 후손 큐보이드를 연합시킨다면 선조 큐보이드 〈2,1,1〉를 추론할 수 있게 된다. 왜냐하면 선조 큐보이드의 완벽한 후손은 3개의 큐보이드로 구성되기 때문이다. 즉, 타킷 큐보이드 〈2,1,1〉를 추론하기 위해서는 서로 간에는 선조-후손 관계를 논할 수 없는(non-comparable) 직계 후손 큐보이드 〈2,1,2〉, 〈2,2,1〉, 〈3,1,1〉을 취급하면 된다. 이는 Basis로서 나타내며, 다음과 같다.

Basis(S, 〈2,1,1〉) = { 〈2,1,2〉, 〈2,2,1〉, 〈3,1,1〉 }

여기에서 큐보이드 〈2,1,1〉을 추론할 수 있는 S 큐보이드는 3개의 직계 후손 큐보이드뿐만 아니라, 이들 3개 큐보이드의 모든 후손 큐보이드들도 해당된다. 그러나 그림 7과 같은 격자 구조의 데이터 큐브에서는 선조의 추론에 필요한 최대 정보를 직계 후손에서 모두 얻을 수 있고, 또한 직계 후손의 모든 후손들로부터 얻을 수 있는 정보는 직계 후손들로부터 얻게 되는 정보와 중복(redundant)되기 때문에 큐보이드 〈2,1,1〉을 추론하기 위해서는 직계 후손 큐보이드 〈2,1,2〉, 〈2,2,1〉, 〈3,1,1〉만을 고려하면 된다. 즉, 전술한 Basis()만 고려하면 된다는 뜻이다. 큐보이드 〈2,1,1〉의 추론에 관한 이상의 내용을 본 연구에서 목적하는 큐보이드 〈2,1,1〉의 추론 방지의 내용으로 전환시키면 다음과 같다. 큐보이드 〈2,1,1〉의 추론을 완벽하게 방지시키기 위해서는 Basis(S, 〈2,1,1〉)을 구성하는 3개의 큐보이드중 한 개의 큐보이드만 공개되도록 하여야 한다.

〔6단계〕 공표 가능한 S의 최대 부분집합

그림 7에서 실선 ①의 윗부분에 있는 큐보이드 집합, 즉, S에 속하는 모든 큐보이드를 다 공표하면 타깃의 큐보이드들이 추론되게 된다. 따라서 공표를 하더라도 타깃을 위태롭게 만들지 않는 즉 누설시키지 않는 S의 최대 부분집합을 구하는 것이 중요하게 된다.

S의 최대 부분집합을 구하자면 먼저 루트(root)를 정의해야 한다. 여기서 루트가 될 수 있는 큐보이드는 Basis()의 구성 큐보이드중 하나만을 자기의 직계 후손 큐보이드로 갖는 타깃의 직계 후손 큐보이드중 하나가 된다. 본 예에서는 큐보이드 〈1,1,2〉, 〈1,2,1〉이 여기에 해당된다. 다음으로 S중 자신의 모든 후손집합이 최대가 되도록 하는 큐보이드를 선택하면 된다. 본 예에서는 큐보이드 〈1,1,2〉가 루트로 선정되며, 이 루트의 모든 후손집합이 S의 최대 부분집합이 된다. 그림 7에서 실선 ②의 윗부분에 있는 모든 큐보이드들이 이에 해당된다.

최대한의 정보를 공개하기 위해서는 그림 7의 실선 ①의 윗부분에 있는 모든 데이터를 공표하는 것이 최선이지만, 이 경우 중요데이터의 직계 후손 큐보이드로부터 중요데이터의 추론이 가능하게 된다. 따라서 추론이 가능한 큐보이드의 공개 제한이 필요하며, 중요 데이터의 추론이 가능한 실선 ②의

아래 부분을 제외한 큐보이드만을 공개해야 한다. 즉, 정보손실을 최소화하고 중요데이터를 보호하면서 공개 가능한 큐보이드는 실선 ②의 윗부분이다.

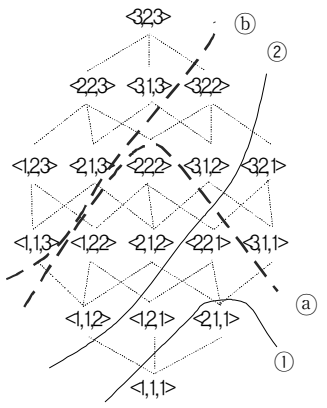


그림 7. 새로운 특허 큐브의 추론통제
Fig. 7. The Inference Control of New Patent Cube

4.2 결과 및 분석

4.1의 추론통제 모델에서 대륙별 각 기관의 평균 특허출원경비, 즉 연도별, 대륙별 기관의 평균 특허출원경비 정보보다 더 개괄적인 정보는 공표해도 되지만 그렇지 않은 경우는 절대적으로 보호해야 한다는 정보보안 방침을 적용하면, 그림 7에서 실선 ②의 윗부분에 있는 큐보이드는 공표 가능한 S의 최대 부분집합임을 알았다. 이하 정보보안 방침의 변화에 따라서 공개 큐보이드가 어떻게 변하는가에 대해 알아보고자 한다.

이를 위하여 기존에는 특허출원한 기관별 정보까지 보호해야 하였지만, 새로운 정보보안 방침은 지역별 정보까지 보호하는 것으로 강화되었다고 하자. 다시 말하면, 연도별, 지역의 특허출원경비 정보보다 더 개괄적인 정보는 공표하고 그렇지 않은 정보는 보호한다는 것으로 볼 수 있다. 이에 따라 보호해야 할 중요 정보 큐보이드는 ⟨Year, ALL, Category⟩ 즉, ⟨2,2,2⟩로 지정할 수 있다. 이 큐보이드의 선조 큐보이드는 ⟨1,2,2⟩, ⟨2,1,2⟩, ⟨2,2,1⟩, ⟨1,1,2⟩, ⟨1,2,1⟩, ⟨2,1,1⟩, ⟨1,1,1⟩로써 이들 7개의 큐보이드는 큐보이드 ⟨2,2,2⟩보다 더 상세한 정보를 가지고 있기 때문에 보호해야 할 큐보이드이며, 그림 7에서 실선 ⓐ의 아래 부분에 해당한다.

Below(⟨2,2,2⟩) = { ⟨2,2,2⟩, ⟨1,2,2⟩, ⟨2,1,2⟩, ⟨2,2,1⟩, ⟨1,1,2⟩, ⟨1,2,1⟩, ⟨2,1,1⟩, ⟨1,1,1⟩ }

다음으로, 타킷 큐보이드 ⟨2,2,2⟩의 직계 후손 큐보이드 ⟨2,2,3⟩과 ⟨3,2,2⟩는 Basis()가 됨을 알 수 있을 것이다.

Basis(S, ⟨2,2,2⟩) = { ⟨2,2,3⟩, ⟨3,2,2⟩ }

데이터 큐브에서 중요데이터가 누설됨을 막기 위해서는 Basis(S, ⟨2,2,2⟩)의 구성 큐보이드 ⟨2,2,3⟩, ⟨3,2,2⟩중 오직 하나의 큐보이드만이 루트의 후손 큐보이드가 되게 하여야 한다.

공표가능 S중 최대 부분집합을 구성시키는 루트를 구해보면 큐보이드 ⟨1,1,3⟩이 되며, 이 큐보이드의 모든 후손집합으로 구성되는 S의 최대 부분집합을 구하면 그림 7에서 실선 ⓑ의 윗부분에 있는 모든 큐보이드들이 된다.

이와 같이 정보보안 방침의 변화, 즉 보호해야 할 중요정보가 기관의 평균 특허출원경비에서 지역별 평균 특허출원경비로 변화함에 따라 그림 7의 실선 ②의 윗부분에서 실선 ⓑ의 윗부분으로 축소 변경되었음을 알 수 있었다. 또한 기존 보안 방침에서 공개 가능하였던 큐보이드 ⟨1,1,2⟩, ⟨1,2,2⟩, ⟨2,1,2⟩, ⟨2,2,2⟩, ⟨3,1,2⟩, ⟨3,2,2⟩ 등 총 6개의 큐보이드가 보안 방침의 강화에 따라 공개 불가능하게 되었다.

표 5. 비공개 큐보이드와 공개 큐보이드
Table 5. The Protected and Unprotected Cuboids

정보보안 방침		연도별, 대륙별, 기관별 평균 특허출원경비를 보호	연도별, 대륙별, 지역별 평균 특허출원경비를 보호
보호	중요 정보 (타킷)	⟨2,1,1⟩, ⟨1,1,1⟩	**⟨2,2,2⟩**, **⟨1,2,2⟩**, **⟨2,1,2⟩**, ⟨2,2,1⟩, **⟨1,1,2⟩**, ⟨1,2,1⟩, ⟨2,1,1⟩, ⟨1,1,1⟩
	비공개	⟨1,2,1⟩, ⟨2,2,1⟩, ⟨3,1,1⟩, ⟨3,2,1⟩	⟨3,1,1⟩, **⟨3,1,2⟩**, ⟨3,2,1⟩, **⟨3,2,2⟩**
공개		**⟨1,1,2⟩**, ⟨1,1,3⟩, **⟨1,2,2⟩**, **⟨2,1,2⟩**, ⟨1,2,3⟩, ⟨2,1,3⟩, **⟨2,2,2⟩**, **⟨3,1,2⟩**, ⟨2,2,3⟩, ⟨3,1,3⟩, **⟨3,2,2⟩**, ⟨3,2,3⟩	⟨1,1,3⟩, ⟨1,2,3⟩, ⟨2,1,3⟩, ⟨2,2,3⟩, ⟨3,1,3⟩, ⟨3,2,3⟩

V. 결 론

오늘날 정부에서의 정책수립이나 시행, 그리고 기업의 창업 및 신규사업진출 등 모든 일을 계획함에 있어 근거 및 판단 자료로 가장 많이 활용하는 것 중의 하나는 통계라고 할 수 있다. 특히 정부가 작성하는 통계는 공식통계로서 정부정책의 기본 인프라로 활용되고, 정책입안과 시행에 활용되기 때문에 국민생활에 직접적인 영향을 미치게 된다.

통계를 생성하는 측면과 통계를 활용하는 측면의 기본적인 입장을 살펴보면, 전자는 사업의 추진목적과 보안을 위하여 공표하는 데이터를 축소하려고 하고, 후자는 보다 정확한 근거마련 및 합리적인 판단을 하기 위해 공표되는 데이터의 확

대를 원하고 있다.

우리나라의 경우 중앙정부나 지방정부의 운영에서 과학적인 근거를 바탕으로 정책을 수립하고 집행하는 과학적 행정이 자리를 잡아가고 있고, 기업을 포함한 통계 수요자는 세계시장의 선도 및 연구역량의 제고 등을 위하여 좀 더 구체적인 데이터를 요구하고 있다. 오늘날의 통계 수요자는 집계 데이터를 제공받는데 만족하지 않고, 집계 데이터를 산출하는 소스가 되는 마이크로데이터를 요구하고 있는 것이다.

이 마이크로데이터는 정부측면에서는 효과적인 정책입안과 시행을 하자면 필수적으로 사용해야만 하는 데이터이고, 기업측면에서는 신규사업의 확대와 히트상품 제조 등 연구 및 기획자가 자신의 필요에 맞게 가공할 수 있는 데이터이다. 그러나 마이크로데이터의 제공은 모든 데이터를 제공하는 것으로 개인정보가 공개될 우려가 있고, 개별 데이터의 기밀보호를 선언한 UN의 통계 기밀보호에 관한 6번째 원칙을 위배할 우려가 있으므로, 이의 제공은 매우 제한된 범위 내에서 그리고 엄격한 조건 아래서 이루어져야 한다.

이러한 기밀보호 원칙을 지키자면 마이크로데이터의 활용에 제약이 생기고, 결과적으로 집계 데이터가 더 많이 쓰이게 되며, 집계 데이터를 사용할수록 세밀한 연구를 할 수 없게 된다. 즉, 마이크로데이터를 사용하여 통계정보를 도출할수록 통계정보는 왜곡되지 않으며, 반대로 집계 데이터를 사용하여 통계정보를 도출할수록 통계정보는 왜곡되게 된다. 그러나 마이크로데이터의 제공은 현실적으로 불가능하므로, 될 수 있는 대로 마이크로데이터에 개념적으로 가까운 집계 데이터를 공개하면 UN의 6번째 원칙을 준수하면서도 정부정책 입안과 시행 및 대국민 홍보의 정부 목적을 원활하게 달성할 수 있으리라 본다.

본 연구에서의 활용방안에 대하여 살펴보면 다음과 같다.

첫째, 정부의 연구개발사업의 정보공개는 합동설명회, 전문기관의 사업공고, 각 부처의 통계 및 홍보자료 등에 크게 의존하고 있으며, 기업의 연구개발능력 제고를 위한 추가정보는 공개하지 않고 있다. 또한 정부의 정책입안 및 시행에서도 집계정보만을 제공받기 때문에 세밀한 단계별 전략을 수립하는데 어려움을 겪고 있다. 따라서 국가 연구개발사업에 본 연구의 추론통제 방법을 적용시키면 연구개발사업과 기업보안에 관한 중요정보를 보호함과 동시에 정책입안 및 시행자와 기업 담당자가 단계별 추진전략을 원활히 수립하게끔 할 수 있다.

둘째, 본 연구에서 개발한 추론통제 모델은 계량 통제모델이라는 것이다. 통계 제공자는 기밀보호 때문에 될수록 집계정보를 제공하고자 하고, 통계 수요자는 정확한 전략수립을 위해 될수록 마이크로데이터에 가까운 데이터를 원하고 있는 상황에서 큐보이드의 공개범위를 이론에 근거하여 계량적으로 정할 수 있는 공식통계의 실용적인 추론통제 전략수립 모델로 사용할 수 있다는 뜻이다.

본 연구에서 제안하고 있는 공식통계의 추론통제 모델의 적용분야는 아주 폭 넓을 것으로 생각한다. 특히 정부의 행정업무, 정책목적의 결정, 경제발전 계획수립과 시행 등의 공표자료는 주로 평균값과 횟수로서 공개되고 있는 현실을 감안할 때, 본 연구에서 제안하고 있는 추론통제 방법은 COUNT와 AVG가 많이 응용되는 분야에서 뛰어난 성과를 낼 수 있으리라 생각한다.

참고문헌

〔1〕 Hallgrimur Snorrason, "Applying United Nations Fundamental Principles of Official Statistics," Seminar of the Arab Institute for Training and Research in Statistics, Sept. 2005.

〔2〕 통계개발원, "*국가 통계 제도의 발전*," 통계개발원, 263-292쪽, 2008년 12월.

〔3〕 통계개발원, 〔2〕, 293-308쪽

〔4〕 L. Lakshmanan, J. Pei, and J. Han, "Quotient Cube: How to Summarize the Semantics of a Data Cube," Proceedings of the 28th VLDB Conference, 2002.

〔5〕 L. Brankovic, M. Miller, P. Horak, and G. wrightson, "Usability of Compromise-Free Statistical Databases for Range Sum Queries," Scientific and Statistical database Management, pp. 144-154, 1997.

〔6〕 L. Wang, S. Jajodia, and D. Wijesekera, "Preserving Privacy in On-Line Analytical Processing(OLAP)," Springer, pp. 37-51, 2007.

〔7〕 UN, "Managing Statistical Confidentiality & Microdata Access," UN Economic Commission for Europe Conference of European Statisticians, 2007.

〔8〕 D. Denning and J. Schloerer, "Inference Controls for Statistical Databases," IEEE Computer, Vol. 16, No. 7, pp. 69-82, 1983.

〔9〕 L. Beck, "A Security Mechanism for Statistical Databases," ACM Transactions on Database Systems, Vol. 5, No. 3, pp. 316-338, Sept. 1980.

〔10〕 이승현, 이덕성, 최인수, "OLAP 큐브에서의 집계함수

AVG의 적용," 한국컴퓨터정보학회논문지, 제 14권, 제 1호, 217-228쪽, 2009년 1월.
〔11〕 A. Shoshani, "OLAP and Statistical Database: Similarities and Differences", Principles of Database Systems, pp.185-196, 1997.
〔12〕 Incremental Maintenance for Non-Distributive Aggregate Functions, *http://seminars.di.uoa.gr/infosys/palpanas*
〔13〕 J. Gray et al., "Data Cube: A Relational Algorithm Operator Generalizing Group-By, Cross-Tab, and Sub-Totals," Data Mining and Knowledge Discovery, Vol. 1, pp. 29-53, 1997.
〔14〕 L. Wang, S. Jajodia, and D. Wijesekera, "Securing OLAP Data Cubes Against Privacy Breaches," Proceedings of the 2004 IEEE Symposium in Security and Privacy, 2004.
〔15〕 L. Wang, S. Jajodia, and D. Wijesekera, 〔6〕, pp.131-136.
〔16〕 이덕성, 최인수, "OLAP 데이터 큐브에서의 추론통제 프로세스 설계," 한국컴퓨터정보학회논문지, 제 14권, 제 5호, 183-193쪽, 2009년 5월.
〔17〕 National Science & Technology Information Service, *http://www.ntis.go.kr*
〔18〕 Free Trade Agreement, *http://fta.customs.go.kr*
〔19〕 A. Casali, R. Cicchetti, and L. Lakhal, "Cube Lattices: A Framework for Multidimensional Data Mining," Proceedings of the 3rd SIAM International Conference on Data Mining, SDM, pp. 304-308, 2003.

저 자 소 개

이 덕 성

1990: 전남대학교 산업공학과 공학사
1995: 전남대학교 산업공학과 공학석사
현재 : 숭실대학교 대학원
산업·정보시스템공학과 박사과정
관심분야 : MIS, DW, OLAP, MDX, CRM

최 인 수

1985 : 서울대학교 산업공학과 공학박사
현재 : 숭실대학교 산업·정보시스템공학과 교수
관심분야 : MIS, DW, OLAP, MDX, CRM